REDESIGN OF THE TRUCKING COMMODITY ORIGIN AND DESTINATION SURVEY

François Gagnon and Julie Trépanier, Business Survey Methods Division, Statistics Canada

1. INTRODUCTION

The objective of the Trucking Commodity Origin and Destination Survey (TCOD), which is managed by Statistics Canada's Transportation Division and sponsored by Transport Canada, is to measure the commodity movements and the outputs of the Canadian trucking industry. The estimates produced include total tonnage transported by commodity type, and revenue by origin and destination of the shipments. The estimates are published in *Trucking in Canada*. The survey data is used by federal and provincial governments, trucking associations, members of the industry, universities and research institutions to assess the industry's growth rate and contribution to the Canadian economy and to measure the volume of provincial and inter-provincial trade transported by trucking companies. In addition, the statistics are used by planning boards to help determine the volume of traffic on highways and by trucking companies that are investigating expansion opportunities. The TCOD survey estimates are used as an input to Statistics Canada's System of National Accounts.

Before the redesign, this survey employed a two-stage sample design, where a sample of trucking companies was selected at the first stage, and a sample of shipping documents was selected from each sampled trucking company via personal, on-site visits at the second stage.

For the past four years, the TCOD survey has been undergoing a redesign in order to increase the survey coverage and to improve the quality of the estimates by utilizing a larger sample size and an enhanced methodology. This had to be achieved without increasing operational costs. The first year of production for the new TCOD survey was 2004. The focus of this chapter is to provide a summary of the methodological challenges that were met in the development of this four-stage sample design redesigned survey.

2. BACKGROUND

The previous TCOD survey was developed in the early 1970s. The last minor revision was at the end of the 1980s. In January 2000, Statistics Canada approved a multiyear project, starting in fiscal year 2000-01, to redesign the TCOD survey. The primary goals were to improve the data quality of the survey and to meet the new requirements of data users.

In 2000-2001, the major users of the TCOD survey data participated in consultation sessions and a questionnaire was broadly distributed. Four major groups of users were consulted: Transport Canada, Statistics Canada's System of National Accounts, the provincial governments and the trucking associations. The highlights of this consultation process were:

■ A majority of users recommended that all shipments transported by trucks in Canada should be covered by the redesigned survey. This implied that shipments made by the following companies should be covered:

- All long-distance Canadian trucking companies (only those above $1 million in revenue were covered by the previous TCOD survey)
- All local Canadian trucking companies (not covered by the previous survey);
- All Canadian non-trucking companies involved in some trucking activity, also known as "Private Trucking" (not covered by the previous survey);
- All foreign companies involved in trucking activities in Canada. (These were not covered in the previous survey).

■ The majority of the users asked that the variables collected be the same as the ones collected in the previous TCOD survey, i.e. tonnage transported, distance, commodities carried, transportation revenues generated, and origin and destination of shipments. Some users asked that we collect the value of the commodities transported and information related to inter-modal shipments and transportation of dangerous goods.

■ A majority of the users requested that the estimates by the shipment's origin and destination be produced at the province or territory level, instead of at the region level (groups of provinces) as in the previous TCOD survey.

■ The users would be satisfied with an annual survey compared to the previous TCOD quarterly survey.

■ With regards to the quality of the data, all stakeholders agreed that the precision of the estimates needed to be improved.

Based on the comments received, it was clear that the

key objectives of the redesign should be to increase the survey coverage and to improve the quality of the TCOD survey estimates, especially at detailed levels. An important constraint was that these objectives had to be achieved without any additional production costs. Because of this constraint, the new survey is conducted annually and each step of the survey has been made as cost-efficient as possible.

3. METHODOLOGY OF THE NEW SURVEY

Although the development of the methodology for the redesigned survey was greatly influenced by the users' needs as expressed in the consultation process, the budget constraints prevented us from completely satisfying all of their requests. The rest of the paper describes how the previous TCOD survey was improved to be more cost-efficient and better meet the users' needs.

3.1 Survey Population and Frame

Both the previous and the redesigned TCOD surveys extract their frames from Statistics Canada's Business Register (BR). The BR statistical structure of a business contains from top to bottom four levels of statistical entities: enterprise, company, establishment and location. In the redesigned survey, all shipments made by the companies on the survey frame are in-scope. Shipments of less than 25 kilometers are not excluded as was the case with the previous survey.

Before redesign

The frame consisted of the list of trucking companies on the BR with annual revenues of $1 million or more that were classified as Long-Distance (48412, 48423) or as Used Household and Office Goods Moving (48421) in the North American Industry Classification System (NAICS). Shipments of less than 25 kilometers made by these companies were deemed to be out of scope. The frame was created on January 1st of the reference year. The companies that were birthed on the BR later in the reference year were not considered nor were the non-trucking companies with trucking establishments considered.

After redesign

The new survey population for the first stage consists of all companies on the BR with at least one trucking establishment (NAICS: 484XXX) and at least $1 million in annual revenue. The Local Trucking sector NAICS (48411, 48422) were added to the previous survey coverage. The frame is created on January 1st of the reference year to allow the interviewers to start collection early in the year (see Section 3.3 on Sample Design for more details). However, a sample of births is selected at the end of the reference year from the list of companies that were not in the TCOD survey population on January 1st of the reference year but that appeared in the survey population for at least one day during the reference year.

All shipments made by the companies on the frame are in-scope for the redesigned survey.

Figure 1: Coverage of the previous and the redesigned TCOD surveys (source: Statistics Canada Business Register, 2004).

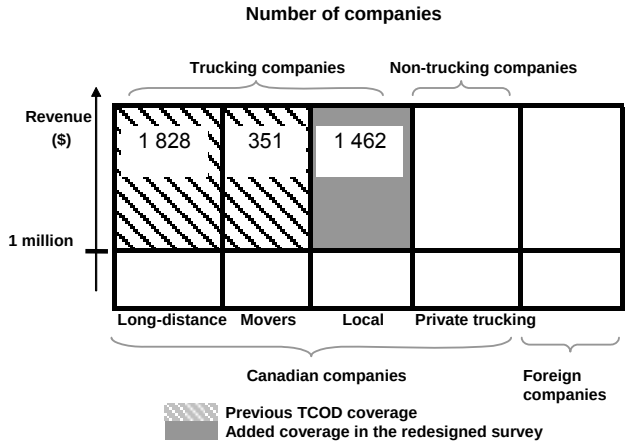

Figure 1 illustrates the change in the survey coverage. In terms of number of companies, there is a 67% increase in the survey coverage [100*1462/(1828+351)]. In terms of revenue, this represents a 35% increase in the survey coverage. Although the survey coverage is significantly enhanced in the redesigned survey, it does not meet all the users' requirements. The Canadian non-trucking companies, also known as "Private Trucking", as well as the foreign companies that transport goods in Canada could not be included in the new survey population due to budget constraints. The $1 million threshold was maintained in the redesigned survey in order to reduce the response burden on small companies and to respect the budget constraints.

3.2 Collection Methods

Both the previous and the redesigned TCOD surveys use three collection methods: (1) electronic data reporting; (2) profiles; and (3) on-site visits.

Before redesign

(1) Electronic Data Reporting. Five companies used to (and continue to) send their data to Statistics Canada using tapes. These tapes contain shipping document information for each shipment made during the reference period. In the previous TCOD survey the coding and imputation processes were not totally automated and only about 10% of the shipments information received electronically was used in the survey.

(2) Profiles. This collection method was used for about 9% of the sampled companies. Statistics Canada interviewers would visit each trucking company selected in the first-stage sample. However, when the trucking company was specialized in some types of shipments (same origin/destination, same commodity, same weight, etc.), the interviewer only collected information about "typical shipments" and noted the number of each "typical shipment" that was made by the trucking company. At the processing stage, synthetic records were generated using this information.

(3) On-site visits. This collection method was used for 91% of the sampled companies in the previous TCOD survey. Statistics Canada interviewers visited each trucking company selected in the sample, selected a sample of these shipping documents and transcribed the data from the sampled documents onto laptop computers. The data was then encrypted and sent by the interviewer to Statistics Canada via modem. This collection process was expensive due to the length of time required for each visit (usually less than a day, but sometimes more), as well as the cost of traveling to and from the company. Furthermore, it was burdensome for the trucking company to have Statistics Canada staff working in their office for long periods of time.

After redesign

The redesigned TCOD survey uses the same three collection methods: (1) electronic data reporting; (2) profiles, but via computer-assisted telephone interviews (CATI); and (3) on-site visits.

(1) Electronic Data Reporting (EDR). For the first year of the new survey (reference year 2004), five companies (the same as in the previous survey) will send us their data electronically. However, 100% of their data will be processed (compared to 10% in the previous survey), thanks to the now totally automated coding and imputation systems. This is expected to add 2 million shipments to the sample. The goal is to significantly increase the number of trucking companies that report in an electronic format so that it becomes, in the long term, the primary collection method for the redesigned survey. This method of collection has the potential to reduce costs, thus permitting us to increase the overall size of the sample of shipments in later stages. It has the potential to improve timeliness (through elimination of the time required to set up interviews, travel and visits), improve response rates (through the reduction in response burden), and improve data quality (e.g., through elimination of data capture errors).

(2) Profiles via CATI. Profiles are used to a greater extent in the redesigned TCOD survey. All companies specialized in some specific types of shipments, i.e. those that reported less than 50 origin/destination/commodity combinations in the previous year, will be collected through a profile via CATI. Instead of visiting the storage location of shipping documents of a company to collect data, the interviewer will collect, through a CATI interview, information about each "typical shipment" and note the number of each "typical shipment" that was made by the company during the reference period. At the processing stage, synthetic records will be generated using this information. The newly developed CATI application will significantly reduce the cost of data collection for those companies whose activities can be described through profiles. Since many more companies will be "profiled" in the redesigned survey we expect to add 2.3 million shipments in the sample.

(3) On-site visits. This collection method is used when neither of the other methods (Profiles or EDR) can be used for a given company. The on-site visits, although reduced in number compared to the previous TCOD survey, remain the most frequent mode of collection in the redesigned survey. As in the case of the previous survey, Statistics Canada interviewers visit each company selected in the sample, select a systematic sample of shipping documents, then select a sample of shipments on each shipping document selected and finally transcribe the data from the documents onto laptop computers. The variables collected for each shipment include: the origin and the destination of the shipment, the description of the commodity transported, the weight and the revenue generated by the shipment. Once transcribed, the data is encrypted and sent by the interviewer to Statistics Canada via modem.

3.3 Sample Design

Before redesign

The previous TCOD survey, a quarterly survey, employed a two-stage sample design where a sample of trucking companies was selected at the first stage, and a systematic sample of shipping documents was selected at the second stage. At the first stage, companies in the survey frame were stratified by their type of operation, area of domicile and annual revenue class (Class I – $12 million and over; Class II – $1 to $12 million). The Class I companies were assigned to take-all strata. The smaller companies were divided into four equal parts within each stratum. Each part represented the sample for one of the four quarters. Consequently, each quarterly sample was a stratified simple random sample selected without replacement.

At the second stage, in order to select a systematic sample of shipments, the interviewer asked the trucking company for a good estimate of the total number of shipments made during the reference period. Using a custom-built software on a laptop computer, the interviewer then used this estimated total number of shipments to find the appropriate sampling interval K that would produce the desired second-stage sample size for this trucking company. Using this sampling interval K and a random start P, a systematic sample of size m shipments was drawn by selecting the shipments P, $P+K$, $P+2K$, …, $P+(m-1)K$. It is important to note that when a company had shipping documents with information about more than one shipment on them, the interviewers had to take out each shipping document and count the number of shipments since the sampling interval was applied at the shipment level, not at the shipping document level. This was time-consuming for the interviewers.

After redesign

The redesigned survey is an annual survey that employs a four-stage sample design where a stratified simple random sample of trucking companies is selected without replacement at the first stage. Companies in the survey frame that can provide their data electronically are part of "EDR" must-take strata. Similarly, the companies that reported less than 50 origin/destination/commodity combinations in the previous year (and that will, therefore, be collected through a profile via CATI) are part of "Profiles" must-take strata. All companies in the must-take strata are in the sample. The companies that are not part of any of the must-take strata are then stratified by three NAICS sectors (long-distance, local and moving companies), 13 activity types based on historical data (for example, transportation of forest products, dry bulk, etc.), 13 provinces or territories of domicile and two annual revenue classes (Class I: larger companies; Class II – smaller companies). The larger companies are part of take-all strata and the smaller companies are part of take-some strata. The Lavallée-Hidiroglou method (1988) is used to optimally determine, in each cell, the threshold between take-all and take-some revenue strata.

At the second stage, for each company selected in the first-stage sample, a period of time (e.g., January to June) is randomly selected within the reference year. This allows spreading the collection workload over the year. For the 2004 reference year, three periods of time are possible. The companies to be collected with a CATI profile, as well as those that provide their data electronically, are assigned a 12-month period of time (January to December). The larger take-all companies are assigned a 12-month collection period as well, while each of the smaller take-all or take-some companies are randomly assigned a 6-month period of time, either the January to June, or the July to December period.

In the redesigned survey, the fixed budget for collection is an important constraint. In the previous survey, some problems occurred when the total collection budget was exhausted before all companies in the sample were collected. Knowing that the remote companies are often the last ones to be visited, the problem of not being able to collect all companies resulted in a possible bias. To avoid such a situation in the redesigned survey, an algorithm was developed to estimate collection costs at the time of sample selection in order to make sure that the collection budget will be sufficient to collect all companies in the sample. The algorithm uses historical information on collection costs associated with the on-site visits (e.g., travel costs to the company location and costs associated with the transcription of shipment data) to ensure that the resulting sample of shipments is likely to be collected within the total budget available. However, the redesigned TCOD survey remains vulnerable as fluctuations in data collection costs cannot be totally predicted (e.g., travel costs, airfare, etc.).

At the third stage, all shipments of companies in EDR or Profiles must-take strata are selected. For the rest of the sampled companies a systematic sample of shipping documents is selected via personal, on-site visits, from each sampled trucking company for the given second-stage period of time. To do so, the interviewer must get a good estimate R_{it} of the total number of shipments transported by the company i during the period of time t. Assuming that there is one shipment per shipping document, custom-built software

on a laptop computer then uses this estimated total number of shipments to derive the appropriate sampling interval A that will give the desired third-stage sample size for this company. Using this sampling interval A and a random start P, a systematic sample (of approximate size r_{it}) is drawn by selecting the shipping documents P, $P+A$, $P+2A$, …, $P+(r_{it}-1)A$. If $R_{it} < 100$ then the sampling interval A is equal to "1". If $R_{it} \geq 100$ then the sampling interval A is given by $ROUND\ (R_{it} / r_{it})$ where r_{it} = round $[11*(R_{it} + 1100)^{0,31}]$. This continuous function replaces the non-continuous function that was used in the previous TCOD to determine the sampling interval and the sample size. This new continuous function gives approximately the same sample size as the old function for a given R_{it} but has the advantage of being a strictly increasing function.

When there is more than one shipment on the shipping document selected at the third stage, a fourth-stage sampling process is involved, in which a systematic sample of shipments is selected from the shipping document. The sampling interval B is given by $B = round\ (M_{itj} / m_{itj})$ where: M_{itj} is the number of shipments on the shipping document j of company i for the period of time t (to be provided by the interviewer to the laptop application); $m_{itj} = 1$ if $M_{itj} = 1$; $m_{itj} = 2$ if $2 \leq M_{itj} \leq 10$; $m_{itj} = 3$ if $11 \leq M_{itj} \leq 30$; and $m_{itj} = 4$ if $M_{itj} \geq 31$. Shipping documents containing multiple shipments are becoming more prevalent (e.g., a monthly invoice to a client that includes information on all shipments made that month for that client). The fourth sampling stage was added to reduce the interviewers' workload when there are many shipments per shipping document. With the addition of this 4th sampling stage, they only have to count the number of shipments on the shipping documents selected at the third sampling stage (in the previous survey, the interviewers had to count the shipments on all shipping documents since the sampling interval was applied at the shipment level only).

3.4 Data Processing

In the redesigned TCOD survey, automation of the coding and imputation processes was a key enhancement to permit the processing of all EDR shipments. The variables collected for each shipment include: the origin and the destination of the shipment, the description of the commodity transported, the weight and the revenue generated by the shipment. The coding of the commodity descriptions, the origin and the destination are done by the interviewers for the companies that are collected by on-site visits or via CATI. To code the commodity description, the interviewer chooses from a list of possible descriptions available in a commodity library on the laptop application. The coding process for the origin and destination is done in a similar manner; the interviewer chooses from a list of possible cities, towns and villages.

The coding process is more complex for the companies that provide their data electronically. The information provided by each company has to be automatically converted into a standard format and nomenclature. This requires us to build mapping or link tables for each company. In the previous survey, the coding process for electronic data was not automated; therefore, we were limited to a sample of about 10% of the electronic data received. In the redesigned TCOD survey, all commodity descriptions obtained in electronic format are automatically coded using Statistics Canada's Automated Coding Text Recognition generalized system. Another automated coding process is used to code the origin and destination variables and generate distances traveled. Moreover, in the redesigned survey, the imputation process has been improved and fully automated.

3.5 Estimation

Before redesign

For the reference period, the estimates of totals for domains of interest were produced using a simple expansion estimator for a two-stage sample design:

$$\hat{Y}(d) = \sum_{h=1}^{H} \sum_{i=1}^{n_h} w_{hi} \sum_{j=1}^{m_{hi}} w_{hij}\, y_{hij}(d)$$

where:

- H = number of strata at the first stage
- n_h = number of companies in the first-stage sample in stratum h
- m_{hi} = number of shipments of company i in stratum h, selected in the 2nd-stage sample
- w_{hi} = first-stage design weight of company i in stratum h
- w_{hij} = second-stage design weight of shipment j of company i in stratum h
- y_{hij} = value of the variable of interest for shipment j of company i in stratum h
- d = domain of interest

$y_{hij}(d) = y_{hij}$ if $hij \in d$, $y_{hij}(d) = 0$ if $hij \notin d$.

The main domain estimates can be described as a 3-dimensional table. In the previous survey, the first two dimensions were the region of origin (5) and the region of destination (5) of the shipment, respectively. The third dimension was the type of commodity (60). As a result of the chosen sample design, the sample size

within each cell of the 3-dimensional table was random. Annual and quarterly estimates were produced.

The variance had two components: the first-stage component took into account the variability between companies while the second-stage component took into account the variability between shipments within a company. These two components were estimated using a Horvitz-Thompson variance estimator for a two-stage design using stratified SRSWOR at the first stage and SRSWOR at the second stage. The variance of the estimator used to produce estimates for a quarter *q* had the following form:

$\mathrm{Var}_q = \mathrm{Var}_q$ (inter-companies) + Var_q (inter-shipments).

The variance of the estimator used to produce annual estimates had the following form:

$\mathrm{Var} = \sum_{q=1 \text{ to } 4} \mathrm{Var}_q + 2 \sum_{q=1 \text{ to } 4} \sum_{q<r} \mathrm{COV}_{q,r}$.

The covariance term was essential since the quarterly samples were not independent within a given reference year.

After redesign

For the reference period, the estimates of totals for domains of interest will be produced using a simple expansion estimator for a four-stage sample design:

$$\hat{Y}(d) = \sum_{h=1}^{H} \sum_{i=1}^{n_h} w_{1hi} w_{2hit} \sum_{j=1}^{r_{hit}} w_{3hitj} \sum_{k=1}^{m_{hitj}} w_{4hitjk} y_{hitjk}(d)$$

where:

H = number of strata at the first stage

n_h = number of companies in the first-stage sample in stratum h

r_{hit} = number of sampled shipping documents, for survey period t , company i in stratum h

m_{hitj} = number of sampled shipments on shipping document j, for survey period t , company i in stratum h

w_{1hi} = first-stage design weight of company i in stratum h

w_{2hit} = 2nd-stage design weight of survey period t for the company i in stratum h

w_{3hitj} = 3rd-stage design weight of shipping document j for period t of company i in stratum h

w_{4hitjk} = 4th stage design weight of shipment k, on shipping document j for period t of company i in stratum h

y_{hitjk} = value of the variable of interest for the shipment k on shipping document j from the survey period t of company i in stratum h

d = domain of interest $y_{hitjk}(d) = y_{hitjk}$ if $hitjk \in d$, $y_{hitjk}(d) = 0$ if $hitjk \notin d$.

Again, the main domain estimates can be described as a 3-dimensional table. The first two dimensions will be the province/territory of origin (13) and the province/territory of destination (13) of the shipment, respectively. The third dimension is the type of commodity (60). As a result of the chosen sample design, the sample size within each cell of the 3-dimensional table is still random. The overall sample size of shipments is expected to increase from 500,000 to approximately 4.9 million in the redesigned survey.. This will improve the quality of the estimates. However, we may still face situations where some domains have too few sampled shipments which would impact on the quality of the estimates.

Under a special data-sharing agreement, Statistics Canada provides Transport Canada, a sponsor of the survey, with a microdata file. As a result, it was important to choose a variance estimation method that would provide Transport Canada with a convenient way of producing variance estimates for the ad-hoc domain estimates that they will generate from the microdata file. We first considered the bootstrap method for the calculation of the variance estimates in the redesigned survey. This method consists of sub-sampling the initial sample. Within each stratum, a simple random sample (SRS) of *n-1* clusters (i.e. first-stage units = companies) is selected with replacement, from the *n* clusters of the stratum. This creates *B* new samples (or repetitions). The same estimate is then calculated for each of the *B* samples, which gives *B* different estimates $\hat{\theta}^*_{(1)}, \hat{\theta}^*_{(2)}, \ldots, \hat{\theta}^*_{(B)}$. To obtain each of the *B* estimates, a specific weight for each sample is necessary. In each SRS sample, the weight is then recalculated for each record in the stratum. These *B* weights, the bootstrap weights, have to be produced and are added to the microdata file. Then, the variance estimator for $\hat{\theta}$ is given by: $v_B(\hat{\theta}) = \frac{1}{B} \sum_b (\hat{\theta}^*_{(b)} - \hat{\theta}^*_{(\bullet)})^2$ where $\theta_{(\bullet)}^* = \frac{\sum_n \hat{\theta}^*_{(b)}}{B}$. Preliminary tests have shown that, in the context of the TCOD survey, the bootstrap variance estimator overestimates the true sampling variance. This is due to the high first-stage sampling fractions in some of the strata. Due to this problem, it was decided to use a Horvitz-Thompson variance estimator for a four-stage sample design instead and to develop a user-friendly module that would allow

Transport Canada to estimate the variance of each of the estimates they produce from the microdata file they receive.

4. CONCLUSION

Most of the objectives of the TCOD redesign have been achieved. The coverage has been increased significantly in terms of both revenue and number of companies, and a more efficient stratification and allocation at the first stage has been implemented. The collection methods have been improved: more data is now collected by more cost-efficient methods such as CATI and EDR. These collection changes have allowed us to increase the expected sample size of shipments from 0.5 M to 4.9 M while keeping the collection costs at the same level as the previous TCOD survey. The coding and imputation of the commodity, origin and destination descriptions have been fully automated, allowing us to increase significantly the quantity of electronic data that can be processed. A new user friendly variance estimation system is being developed to allow Transport Canada to produce their own variance estimates from the microdata file they receive. Another goal is to continue to increase the number of companies that report their data electronically in order to both increase the sample size of shipments and reduce the collection costs.

ACKNOWLEDGEMENTS

The authors would like to thank Sébastien Landry, Jeannine Claveau, Jean-François Bastien and Windie Gagné for their valuable contribution to this project. The authors would also like to thank Ed Hamilton and Jack Lothian for their comments that helped to improve the quality of the paper.

REFERENCES

Gagnon, F., Rathwell, S. and Gauthier, Y. (2000). Electronic Data Reporting in the Context of the Redesign of the Canadian For-Hire Trucking Origin/Destination Survey. *Proceedings of the Second International Conference on Establishment Surveys*, American Statistical Association, 171-179.

Lavallée, P., and Hidiroglou, M.A. (1988), On the Stratification of Skewed Populations. *Survey Methodology*, Volume 14, Number 1, 33-43.

Rao, J.N.K., Wu, C.F.J., and Yue, K. (1992). Some Recent Work on Resampling Methods for Complex Surveys. *Survey Methodology*, Volume 18, Number 2, 209-217.

Trucking Traffic Survey Development Project Team (2001), Data Needs and Requirements Document: Findings of the Consultation with Major Stakeholders. *Unpublished report*, Statistics Canada.